

WE CLAIM:

1. A synthetic gene encoding a polypeptide segment that corresponds to a reference polypeptide segment encoded by a naturally occurring gene, wherein the polypeptide segment-encoding sequence of the synthetic gene is different from the polypeptide segment-encoding sequence of said naturally occurring gene, wherein

a) said polypeptide segment-encoding sequence of said synthetic gene is less than about 90% identical to said polypeptide segment-encoding sequence of said naturally occurring gene, and/or

b) said polypeptide segment-encoding sequence of said synthetic gene comprises at least one unique restriction site that is not present or is not unique in the polypeptide segment-encoding sequence of said naturally occurring gene, and/or

c) said polypeptide segment-encoding sequence of said synthetic gene is free from at least one restriction site that is present in the polypeptide segment-encoding sequence of said naturally occurring gene.

2. The synthetic gene of claim 1 wherein the polypeptide segment is from a polyketide synthase (PKS).

3. The synthetic gene of claim 2 wherein the polypeptide segment comprises a PKS domain selected from AT, ACP, KS, KR, DH, ER, and TE.

4. The synthetic gene of claim 3 that encodes one or more PKS modules.

5. The synthetic gene of claim 4 comprising at most one copy per module-encoding sequence of a restriction enzyme recognition site selected from the group consisting of Spe I, Mfe I, Afi II, Bsi WI, Sac II, Ngo MIV, Nhe I, Kpn I, Msc I, Bgl II, Bss HII, Sac II, Age I, Pst I, Kas I, Mlu I, Xba I, Sph I, Bsp E, and Ngo MIV recognition sites.

6. The synthetic gene of claim 1 wherein the polypeptide segment-encoding sequence of the synthetic gene is free from at least one Type IIS enzyme restriction site present in the polypeptide segment-encoding sequence of said naturally occurring gene.
7. A synthetic gene encoding a polypeptide segment that corresponds to a reference polypeptide segment encoded by a naturally occurring PKS gene, wherein the polypeptide segment-encoding sequence of the synthetic gene is different from the polypeptide segment-encoding sequence of said naturally occurring PKS gene and comprises at least two of:
 - a) a Spe I site near the sequence encoding the amino-terminus of the module;
 - b) a Mfe I site near the sequence encoding the amino-terminus of a KS domain;
 - c) a Kpn I site near the sequence encoding the carboxy-terminus of a KS domain;
 - d) a Msc I site near the sequence encoding the amino-terminus of an AT domain;
 - e) a Pst I site near the sequence encoding the carboxy-terminus of an AT domain;
 - f) a BsrB I site near the sequence encoding the amino-terminus of an ER domain;
 - g) an Age I site near the sequence encoding the amino-terminus of a KR domain;
 - h) an Xba I site near the sequence encoding the amino-terminus of an ACP domain.
8. A vector comprising a synthetic gene of claim 1.
9. The vector of claim 8 that is an expression vector.
10. A library of vectors each comprising a synthetic gene of claim 1.
11. The vector of claim 8 that comprises an open reading frame encoding a first PKS module and one or more of:
 - a) a PKS extension module;
 - b) a PKS loading module;
 - c) a thioesterase domain; and
 - d) an interpolypeptide linker.

12. A cell comprising an expression vector of claim 9.
13. The cell of claim 12 comprising a polypeptide encoded by the vector.
14. The cell of claim 13 that comprises a functional polyketide synthase, wherein said PKS comprises a polypeptide encoded by said vector.
15. A method of making a polyketide comprising culturing a cell of claim 14 under conditions in which a polyketide is produced, wherein the polyketide would not be produced by said cell in the absence of said vector.
16. A gene library comprising a plurality of different PKS module-encoding genes, wherein the module-encoding genes in the library have at least one restriction site in common, said restriction site is found no more than one time in each module, and the modules encoded in said library correspond to modules from five or more different polyketide synthase proteins.
17. The library of claim 16 wherein said module-encoding genes comprise at least three restriction sites in common.
18. The library of claim 16 wherein the unique restriction is selected from the group consisting of consisting of Spe I, Mfe I, Afi II, Bsi WI, Sac II, Ngo MIV, Nhe I, Kpn I, Msc I, Bgl II, Bss HII, Sac II, Age I, Pst I, Bsr BI, Kas I, Mlu I, Xba I, Sph I, Bsp E, and Ngo MIV recognition sites.
19. The library of claim 16 wherein said at least one restriction site in common is:
a) a Spe I site near the sequence encoding the amino-termini of the modules;
and/or
b) a Mfe I site near the sequence encoding the amino-termini of KS domains;
and/or
c) a Kpn I site near the sequence encoding the carboxy-termini of KS domains;
and/or

- d) a Msc I site near the sequence encoding the amino-termini of AT domains;
and/or
- e) a Pst I site near the sequence encoding the carboxy-termini of AT domains;
and/or
- f) a BsrB I site near the sequence encoding the amino-termini of ER domains;
and/or
- g) a Age I site near the sequence encoding the amino-termini of KR domains;
and/or
- h) a Xba I site near the sequence encoding the amino-termini of ACP domains.

20. The library of claims 16 wherein said genes are contained in cloning or expression vectors.

21. The library of claim 20 wherein each PKS module-encoding gene also comprises coding sequence for

- a) at least a second PKS extension module, or
- b) a PKS loading module, or
- c) a thioesterase domain, or
- d) an interpolypeptide linker.

22. A cloning vector comprising, in the order shown,

- a) SM4 – SIS – SM2 – R₁ *or*
- b) L – SIS – SM2 – R₁

where SIS is a synthon insertion site, SM2 is a sequence encoding a first selectable marker, SM4 is a sequence encoding a second selectable marker different from the first, R₁ is a recognition site for a restriction enzyme, and L is a recognition site for a different restriction enzyme.

23. A vector of claim 22 wherein SM2 and SM4 are genes conferring drug resistance.

24. A composition comprising a vector of claim 1 and a restriction enzyme that recognizes R_1 .
25. The cloning vector of claim 22 wherein the SIS comprises $-N_1-R_2-N_2-$ where N_1 and N_2 are recognition sites for nicking enzymes, and may be the same or different, and R_2 is a recognition site for a restriction enzyme different from R_1 or L .
26. A composition comprising a vector of claim 25 and a nicking enzyme.
27. A vector comprising
- a) $SM4 - 2S_1 - Sy_1 - 2S_2 - SM2 - R_1$ or
 - b) $L - 2S_1 - Sy_2 - 2S_2 - SM2 - R_1$
- where $2S_1$ is a recognition site for first Type IIS restriction enzyme,
 where $2S_2$ is a recognition site for a different Type IIS restriction enzyme, and Sy is synthon coding region.
28. The vector of claim 27 wherein Sy encodes a polypeptide segment of a polyketide synthase.
29. A composition comprising a vector of claim 26 and a Type IIS restriction enzyme that recognizes either $2S_1$ or $2S_2$.
30. A composition comprising a cognate pair of vectors, wherein said cognate pairs are:
- a) a first vector comprising $SM4-2S_1-Sy_1-2S_2-SM2-R_1$ digested with a Type IIS restriction enzyme that recognizes $2S_2$, and
 a second vector comprising $SM5-2S_3-Sy_2-2S_4-SM3-R_1$ digested with a Type IIS restriction enzyme that recognizes $2S_3$;
 - or
 - b) a first vector comprising $L-2S_1-Sy_1-2S_2-SM2-R_1$ digested with a Type IIS restriction enzyme that recognizes $2S_2$, and

a second vector comprising L'-2S₃-Sy₂-2S₄-SM3-R₁ digested with a Type IIS restriction enzyme that recognizes 2S₃,
wherein SM1, SM2, SM3, SM4 are sequences encoding different selection markers, R₁ is a recognition site for a restriction enzyme, L and L' are recognition sites that are the same or the same or different, and each different from R₁, 2S₁, 2S₂, 2S₃, and 2S₄ are recognition sites for Type IIS restriction enzymes, wherein 2S₁, 2S₂ are not the same, 2S₃, and 2S₄ are not the same, and digestion of the first vector with 2S₂ and the second vector with 2S₃ results in compatible ends.

31. The composition of claim 30 wherein 2S₁ and 2S₃ are the same and 2S₂ and 2S₄ are the same.

32. The composition of claim 30 wherein Sy₁ and Sy₂ encode polypeptide segments of a polyketide synthase.

33. A vector comprising a first selectable marker, a restriction site (R₁) recognized by a first restriction enzyme, and a synthon coding region flanked by a restriction site recognized by a first Type IIS restriction enzyme and a restriction site recognized by a second Type IIS restriction enzyme

wherein digestion of the vector with said first restriction enzyme and said first Type IIS restriction enzyme produces a fragment comprising said first selectable marker and said synthon coding region, and

digestion of the vector with said first restriction enzyme and said second Type IIS restriction enzyme produces a fragment comprising said synthon coding region and not comprising said first selectable marker.

34. A method for joining a series of DNA units using a vector pair comprising
a) providing a first set of DNA units, each in a first-type selectable vector comprising a first selectable marker and providing a second set of DNA units, each in a second-type selectable vector comprising a second selectable marker different from the first,

wherein said first-type and second-type selectable vectors can be selected based on the different selectable markers,

b) recombinantly joining a DNA unit from the first set with an adjacent DNA unit from the second set to generate a first-type selectable vector comprising a third DNA unit, and obtaining a desired clone by selecting for the first selectable marker

c) recombinantly joining the third DNA unit with an adjacent DNA unit from the second set to generate a first-type selectable vector comprising a fourth DNA unit, and obtaining a desired clone by selecting for the first selectable marker, or

recombinantly joining the third DNA unit with an adjacent DNA unit from the second series to generate a second-type selectable vector comprising a fourth DNA unit, and obtaining a desired clone by selecting for the second selectable marker.

35. The method of claim 34 wherein step (c) comprises recombinantly joining the third DNA unit with an adjacent DNA unit from the second set to generate a first-type selectable vector comprising a fourth DNA unit, and obtaining a desired clone by selecting for the first selectable marker, said method further comprising

recombinantly combining the fourth DNA unit with an adjacent DNA unit from the second series to generate a first-type selectable vector comprising a fifth DNA unit, and obtaining a desired clone by selecting for the first selection marker, or

recombinantly combining the third DNA unit with an adjacent DNA unit from the second set to generate a second-type selectable vector comprising a fifth DNA unit, and obtaining a desired clone by selecting for the second selection marker.

36. The method of claim 34 wherein step (c) comprises recombinantly joining the third DNA unit with an adjacent DNA unit from the second series to generate a second-type selectable vector comprising a fourth DNA unit, and obtaining a desired clone by selecting for the second selectable marker, said method further comprising

recombinantly joining the fourth DNA unit with an adjacent DNA unit from the first set to generate a first-type selectable vector comprising a fifth DNA unit, and obtaining a desired clone by selecting for the first selection marker, or

recombinantly joining the third DNA unit with an adjacent DNA unit from the first set to generate a second-type selectable vector comprising a fifth DNA unit and obtaining a desired clone by selecting for the second selection marker.

37. The method of claim 34 wherein the desired clone comprises a sequence encoding a PKS domain.

38. A method for joining several DNA units in sequence, said method comprising
a) carrying out a first round of stitching comprising ligating an acceptor vector fragment comprising a first synthon SA_0 , a ligatable end LA_0 at the junction end of synthon SA_0 and an adjacent synthon SD_0 , and another ligatable end la_0 ,

and a donor vector fragment comprising a second synthon SD_0 , a ligatable end LD_0 at the junction end of synthon SD_0 and synthon SA_0 , wherein LD_0 and LA_0 are compatible, another ligatable end ld_0 , wherein ld_0 and la_0 are compatible, and a selectable marker,

wherein LA_0 and LD_0 are ligated and la_0 and ld_0 are ligated, thereby joining said first and second synthons, and thereby generating a first vector comprising synthon coding sequence S_1 ;

b) selecting for said first vector by selecting for the selectable marker in (a);
and,

c) carrying out a number n additional rounds of stitching,
wherein n is an integer from 1 to 20,
wherein S_n is the synthon coding sequence generated by joining synthons in the previous round of stitching, and

wherein each round n of stitching comprises:

1) designating said first or a subsequent vector as either an acceptor vector A_n or a donor vector D_n

2) digesting acceptor vector A_n with restriction enzymes to produce an acceptor vector fragment comprising a synthon coding sequence S_n , a ligatable end LA_n at the junction end of synthon S_n and an adjacent synthon SD_{n+100} , and another ligatable end la_n ; and,

ligating the acceptor vector fragment to a donor vector fragment comprising synthon SD_{n+100} , a ligatable end LD_{n+100} at the junction end of synthon SD_{n+100} and synthon S_n , wherein LA_n and LD_{n+100} are compatible. another ligatable end ld_{n+100} , wherein la_n and ld_{n+100} are compatible, and a selectable marker,

wherein LA_n and LD_{n+100} are ligated and la_n and ld_{n+100} are ligated, thereby generating a subsequent vector, or

digesting donor vector D_n with restriction enzymes to produce a donor vector fragment comprising a synthon coding sequence S_n , a ligatable end LD_n at the junction end of synthon S_n and an adjacent synthon SA_{n+100} , another ligatable end ld_n , and a selectable marker; and

ligating the donor vector fragment to an acceptor vector fragment comprising synthon SA_{n+100} , a ligatable end LA_{n+100} at the junction end of synthon SA_{n+100} and synthon S_n , and another ligatable end la_{n+100}

wherein LA_{n+100} and LD_n are compatible and are ligated and la_{n+100} and ld_n are compatible and are ligated,

thereby generating a subsequent vector

d) selecting the subsequent vector by selecting for the selectable marker of said donor vector fragment of step (c)

e) repeating steps (c) and (d) $n-1$ times thereby producing a multisynthon.

39. The method of claim 1 wherein the selectable marker of step (d) is not the same as the selectable marker of the preceding stitching step and/or is not the same as the selectable marker of the subsequent stitching step.

40. The method of claim 37 wherein la_0 , ld_0 , la_n , ld_n are the same and/or La_0 , Ld_0 , La_n , and Ld_n are created by a Type IIS restriction enzyme.

41. The method of claim 37 wherein said synthons SA_0 , SD_0 , SA_{n+100} , and SD_{n+100} are synthetic DNAs.

42. The method of claim 37 wherein any one or more of synthons SA_0 , SD_0 , SA_{n+100} , or SD_{n+100} is a multisynthon.

43. The method of claim 37 wherein the multisynthon product of step (e) encodes a polypeptide comprising a PKS domain.

44. A method for making a synthetic gene encoding a PKS module, comprising

- (i) producing a plurality of DNA units by assembly PCR, wherein each DNA unit encodes a portion of said PKS module;
- (ii) combining said plurality of DNA units in a predetermined sequence to produce PKS module-encoding gene.

45. The method of claim 44, further comprising combining said module-encoding gene in-frame with a nucleotide sequence encoding a PKS extension module, a PKS loading module, a thioesterase domain, or an PKS interpolypeptide linker, thereby producing a PKS open reading frame.

46. A method for identifying restriction enzyme recognition sites useful for design of synthetic genes, comprising the steps of

- obtaining amino acid sequences for a plurality of functionally related polypeptide segments;
- reverse-translating said amino acid sequences to produce multiple polypeptide segment-encoding nucleic acid sequences for each polypeptide segment;
- identifying restriction enzyme recognition sites that are found in at least one polypeptide segment-encoding nucleic acid sequence of at least about 50% of said polypeptide segments.

47. The method of claim 46 wherein said functionally related polypeptide segments are polyketide synthase modules or domains.

48. The method of claim 46 wherein said functionally related polypeptide segments are regions of high homology in PKS modules or domains.

49. A method for high throughput synthesis of a plurality of different DNA units comprising different polypeptide encoding sequences comprising: for each DNA unit, performing polymerase chain reaction (PCR) amplification of a plurality of overlapping oligonucleotides to generate a DNA unit encoding a polypeptide segment and adding UDG-containing linkers to the 5' and 3' ends of the DNA unit by PCR amplification, thereby generating a linkered DNA unit, wherein the same UDG-containing linkers are added to said different DNA units.

50. The method of claim 49 wherein said plurality comprises more than 50 different DNA units.

51. A method for designing a synthetic gene, the method comprising the steps of:
providing a reference amino acid sequence;
reverse translating the amino acid sequence to a randomized nucleotide sequence which encodes the amino acid sequence using a random selection of codons which have been, optionally, optimized for a codon preference of a host organism;
providing one or more parameters for positions of restriction sites on a sequence of the synthetic gene;
removing occurrences of one or more selected restriction sites from the randomized nucleotide sequence; and
inserting one or more selected restriction sites at selected positions in the randomized nucleotide sequence to generate a sequence of the synthetic gene.

52. The method of claim 51, further comprising:
generating a set of overlapping oligonucleotide sequences which together comprise a sequence of the synthetic gene.

53. The method of claim 54, wherein:
one or more parameters for positions of restriction sites on a sequence of the synthetic gene comprises one or more preselected restriction sites at selected positions.

54. The method of claim 51, wherein the inserting of restriction sites comprises:
identifying selected positions for insertion of a selected restriction site in the randomized nucleotide sequence;
performing a substitution in the nucleotide sequence at the selected position such that the selected restriction site sequence is created at the selected position;
translating the substituted sequence to an amino acid sequence;
accepting a substitution wherein the translated amino acid sequence is identical to the reference amino acid sequence at the selected position and rejecting a substitution wherein the translated amino acid sequence is different from the reference amino acid sequence at the selected position.

55. The method of claim 54, wherein a translated amino acid sequence identical to the reference amino acid sequence comprises substitution of an amino acid with a similar amino acid at the selected position.

56. The method of claim 51, wherein the reference amino acid sequence is of a naturally occurring polypeptide segment.

57. A system for designing a synthetic gene, including a computer processor configured to:
provide a reference amino acid sequence;
reverse translate the amino acid sequence to a randomized nucleotide sequence which encodes the amino acid sequence using a random selection of codons which have been, optionally, optimized for a codon preference of a host organism;
provide one or more parameters for positions of restriction sites on a sequence of the synthetic gene;

remove occurrences of one or more selected restriction sites from the randomized nucleotide sequence;

insert one or more selected restriction sites at selected positions in the randomized nucleotide sequence to generate a sequence of the synthetic gene; and

generate a set of overlapping oligonucleotide sequences which together comprise a sequence of the synthetic gene.

58. A computer readable storage medium containing computer executable code for designing a synthetic gene by instructing a computer to operate as follows:

provide a reference amino acid sequence;

reverse translate the amino acid sequence to a randomized nucleotide sequence which encodes the amino acid sequence using a random selection of codons which have been, optionally, optimized for a codon preference of a host organism;

provide one or more parameters for positions of restriction sites on a sequence of the synthetic gene;

remove occurrences of one or more selected restriction sites from the randomized nucleotide sequence;

insert one or more selected restriction sites at selected positions in the randomized nucleotide sequence to generate a sequence of the synthetic gene; and

generate a set of overlapping oligonucleotide sequences which together comprise a sequence of the synthetic gene.

59. A method for analyzing a nucleotide sequence of a synthon, the method comprising:

providing a sequence of a synthetic gene, wherein the synthetic gene is divided into a plurality of synthons;

providing sequences of a plurality of synthon samples wherein each synthon of the plurality of synthons is cloned in a vector;

providing a sequence of the vector without an insert;

eliminating vector sequences from the sequence of the cloned synthon;

constructing a contig map of sequences of the plurality of synthons;

aligning the contig map of sequences with the sequence of the synthetic gene; and
identifying a measure of alignment for each of the plurality of synthons.

60. The method of claim 59, further comprising:
identifying errors in one or more synthon sequences; and
reporting one or more informations selected from the group consisting of: a
ranking of synthon samples by degree of alignment, an error in the sequence of a synthon
sample, and identity of a synthon that can be repaired.
61. A system for high through-put synthesis of synthetic genes comprising:
at least one source microwell plate containing oligonucleotides for assembly PCR
a source for an assembly PCR amplification mixture
a source for LIC extension primer mixture
at least one PCR microwell plate for amplification of oligonucleotides
a liquid handling device which
retrieves a plurality of predetermined sets of oligonucleotides from the
source microwell plate(s)
combines the predetermined sets and the amplification mixture in wells of
the at least one PCR microwell plate;
retrieves LIC extension primer mixture; and
combines the LIC extension primer mixture and amplicons in a well of the
at least one PCR microwell plate; and
a heat source for PCR amplification configured to accept the at least one PCR
microwell plate.
62. The system of claim 1 further comprising a source for at least two assembly
vectors.
63. An open reading frame vector having a structure selected from
a) Internal type: 4-[7-*]-[*-8]-3;
b) Left-edge type: 4-[7-1]-[*-8]-3; and

c) Right-edge type: 4-[7-*]-[6-8]-3;

wherein 7 and 8 are recognition sites for Type IIS restriction enzymes which cut to produce compatible overhangs “*”; 1 and 6 are Type II restriction enzyme sites that are optionally present; and 3 and 4 are recognition sites for restriction enzymes with 8-basepair recognition sites.

64. The vector of claim 63 wherein 1 is Nde I, 6 is Eco RI, 4 is Not I and 3 is Pac I.